

Conference Paper Title*

Tziporah Horowitz

Johns Hopkins University, Whiting School of Engineering

ACM 725 - Theory of Statistics I

Abstract—This document is a model and instructions for L^AT_EX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.].
***CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

I. INTRODUCTION

Statistical binary classification is a process used in medical testing to determine if a patient has a certain disease or not. It often employs supervised learning to determine decision classes based on predefined rules. While methods such as decision trees, random forests, and neural networks are commonly used in other applications of binary classification, logistic regression offers medical testing a number of advantages due to its probabilistic nature and its relationship with the odds ratio [?]. If interpreted correctly, it can provide a measure of associated risk between the independent variables and the binary outcome.

The upcoming sections will provide an overview of the statistical theory behind logistic regression, followed by an analysis of the association between various risk factors and the chances of a patient contracting heart disease. To measure the associated risk, this paper provides a synopsis of four common methods: *risk difference*, *relative risk*, *odds ratio's*, and *marginal effects*.

II. LOGISTIC REGRESSION

III. METHODS

There are several ways to measure the association between a risk factor and the binary outcome of contracting a disease. The following sections discuss the four approaches used in this analysis to determine the associated risk of various factors in the context of patients contracting heart disease. Sections through refer to table 1 for simplification.

A. Risk Difference

Often considered the simplest approach for measuring associated risk, *risk difference* or *absolute risk difference* (ARD) is the difference in the outcome rates between patients with the risk factor and patients without the risk factor [?]. Using the matrix in 1, risk difference can be defined mathematically as:

$$ARD = \frac{D_E}{D_E + H_E} - \frac{D_U}{D_U + H_U}$$

While the risk difference is easy to compute, its interpretation is often misleading and can only explain the associated risk between a single factor and the target.

B. Relative Risk

Similar to risk difference, *relative risk* compares the outcome rates between patients with the risk factor and patients without the risk factor.

However, relative risk is computed as a ratio (RR) rather than a difference [?]. The risk ratio is defined as:

$$RR = \frac{D_E / (D_E + H_E)}{D_U / (D_U + H_U)}$$

Relative risk is a useful statistic because it quantifies the probability of a patient with exposure contracting the disease relative to a patient without exposure. Risk ratios that are close to 1 indicate that the risk of contracting the disease for an exposed patient is the same as the risk for an unexposed patient. In contrast, risk ratios that are far from 1 indicate that there is an association between the variables. This allows one to create a confidence interval using the hypothesis test,

$$H_0 : RR = 1$$

$$H_1 : RR \neq 1$$

The risk ratio is considered a valid measure of relative risk for studies in which the sampling is dependent on the exposure of interest such as, randomized controlled trials or cohort and cross-sectional studies [?]. Like risk difference, relative risk can only explain the associated risk between a single factor and the target.

C. Odds Ratio

Often confused with risk ratio, *odds ratio* compares the statistical odds of the outcome in the exposed group to that of the outcome of the unexposed group. It is defined mathematically as:

$$OR = \frac{D_E / H_E}{D_U / H_U}$$

TABLE I: Contingency Matrix

	Diseased	Healthy
Exposed	D_E	H_E
Unexposed	D_U	H_U

Like the risk ratio, odds ratios that are close to 1 indicate no association between exposure and contracting the disease, and odds ratios that are far from 1 indicate that there is an association between the variables. One can also create a confidence interval for the odds ratio using a similar hypothesis test to that of the risk ratio, such that

$$H_0 : OR = 1$$

$$H_1 : OR \neq 1$$

While the odds ratio is typically considered the “only valid measure of relative association in traditional case-control studies” [?], it is frequently misinterpreted as the risk ratio. However, in cases where the risk factor is relatively small ($< 10\%$), the odds ratio approximates the risk ratio:

$$\frac{D_E/(D_E + H_E)}{D_U/(D_U + H_U)} \approx \frac{D_E/H_E}{D_U/H_U}$$

if D_E and D_U are small.

The odds ratio can be applied in multi-parameter settings when computed in a logistic regression analysis, due to its inherent calculation of the logit (or log-odds) function. To obtain the odds ratio of a logistic regression model, one simply has to exponentiate the coefficients.

D. Marginal Effects

A *marginal effect* (or incremental effect) is the change in the probability that an outcome occurs as the risk factor changes by one unit. It is often used in logistic regression analysis and other GLM’s

to explain the incremental risk associated with each factor [?]. Marginal effects are determined by taking the partial derivative of the regression equation with respect to each variable. They are simpler to interpret than odds ratios and are easier to compare across different studies. There are many ways to represent the marginal effect for a sample, the most common of which is the *average marginal effect* across all patients in the dataset.

E. The Analysis

For this analysis, the *Heart Disease Dataset* was collected from kaggle.com [?]. The dataset includes data that was compiled from four databases in 1988 and consists of 1025 rows and 14 columns: 13 predictors and 1 target. The predictors include 5 continuous variables: age, resting blood pressure, serum cholestoral (in mg/dl), maximum heart rate achieved, and ST depression induced by exercise relative to rest (oldpeak); and 8 categorical variables: sex, chest pain type, fasting blood sugar > 120 mg/dl (true or false), resting electrocardiographic results, exercise induced angina (yes or no), the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy, and thal (normal, fixed defect, reversible defect). Figure 1 shows the *skimpy* summary of all 14 variables and figure 2 shows the distributions of predictor variables when compared to the target, a binary indicator for the patient having heart disease.

Prior to computing any of the above measures of associated risk, the data was imported into Python














skimpy summary									
Data Summary					Data Types				
dataframe		Values			Column Type		Count		
Number of rows		1025			int32		13		
Number of columns		14			float64		1		
number									
column_name	NA	NA %	mean	sd	p0	p25	p75	p100	hist
age	0	0	54	9.1	29	48	61	77	
sex	0	0	0.7	0.46	0	0	1	1	
cp	0	0	0.94	1	0	0	2	3	
trestbps	0	0	130	18	94	120	140	200	
chol	0	0	250	52	130	210	280	560	
fbs	0	0	0.15	0.36	0	0	0	1	
restecg	0	0	0.53	0.53	0	0	1	2	
thalach	0	0	150	23	71	130	170	200	
exang	0	0	0.34	0.47	0	0	1	1	
oldpeak	0	0	1.1	1.2	0	0	1.8	6.2	
slope	0	0	1.4	0.62	0	1	2	2	
ca	0	0	0.75	1	0	0	1	4	
thal	0	0	2.3	0.62	0	2	3	3	
target	0	0	0.51	0.5	0	0	1	1	
End									

Fig. 1: Summary of Variables

(3.9) using `pandas` and dummy variables were created for the categorical columns. Individual risk differences and risk ratios were then computed for each predictor variable. The dataset was then divided into a 3 : 2 train-test split so that a logistic regression model can be fit. Two models were fit using the training set to the objective function,

$$p = \beta X$$

via `statsmodels`. One was computed with the odds ratio, and the other with marginal effects. The models were then tested with the remainder of the data and scored using `sklearn`.

IV. RESULTS

	Risk Difference	Risk Ratio
age	-0.2892	0.7589
sex	-0.3036	0.7372
chest pain $\in \{1, 2\}$	0.4813	1.6242
resting blood pressure > 130	-0.0888	0.9158
serum cholestorl > 250 ml/dl	-0.1512	0.8594
fasting blood sugar > 120 mg/dl	-0.0577	0.9437
resting electrocardiographic results $\in \{0, 1\}$	0.3178	1.3741
maximum heart rate achieved > 150	0.4067	1.5011
exercise induced angina	-0.4633	0.6287
oldpeak	-0.3773	0.6855
slope	0.1453	1.1563
major vessels colored $\in \{0, 1, 2\}$	0.2593	1.2959
thal $\in \{2, 3\}$	0.1882	1.2074

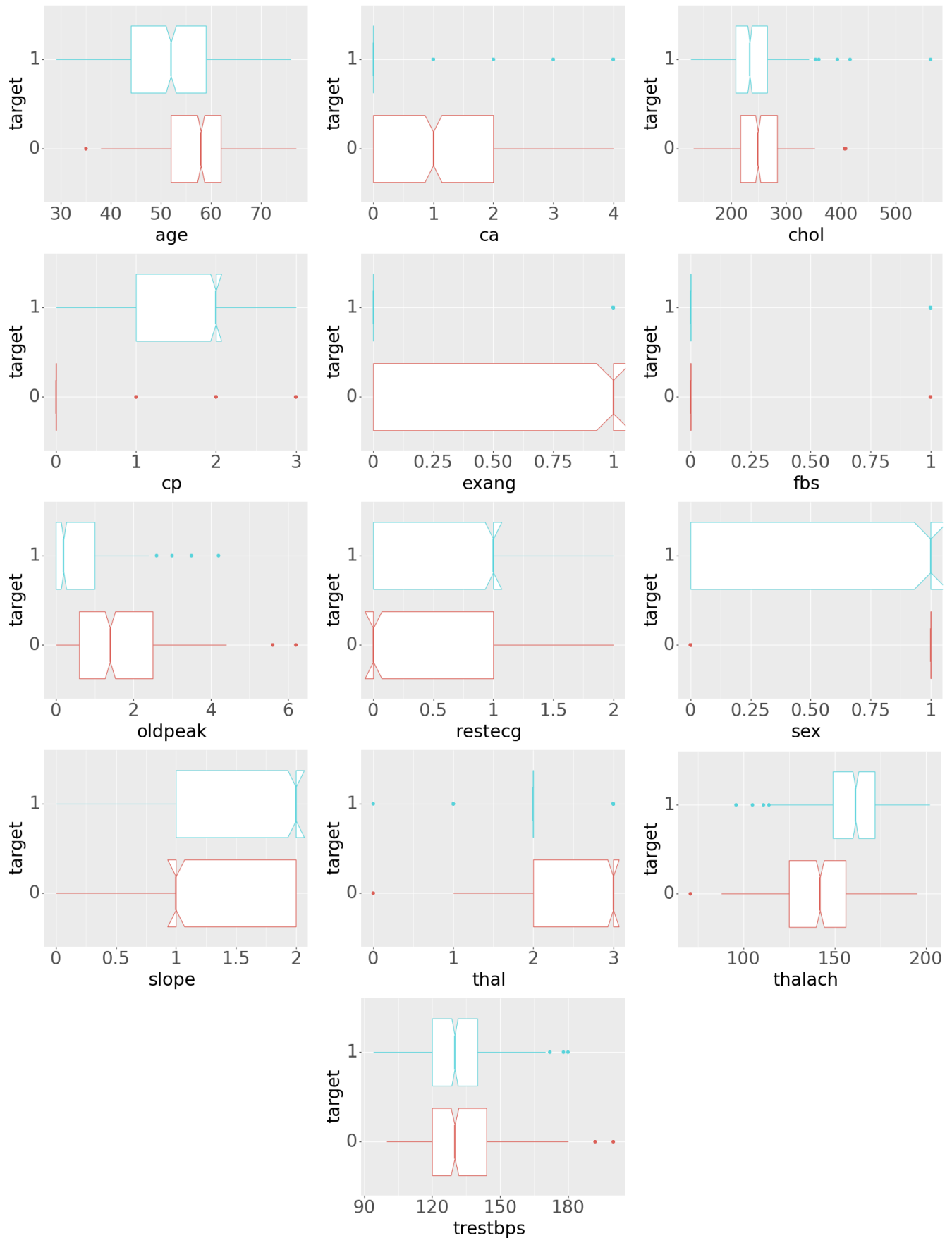


Fig. 2: Distributions of Feature Variables

Dep. Variable:	target	No. Observations:	615			
Model:	Logit	Df Residuals:	602			
Method:	MLE	Df Model:	12			
Date:	Thu, 24 Nov 2022	Pseudo R-squ.:	0.5058			
Time:	15:10:58	Log-Likelihood:	-210.48			
converged:	True	LL-Null:	-425.93			
Covariance Type:	nonrobust	LLR p-value:	1.073e-84			
	coef	std err	z	P> z 	[0.025	0.975]
age	0.0272	0.014	1.983	0.047	0.000	0.054
sex	-1.9158	0.335	-5.712	0.000	-2.573	-1.258
cp	0.8560	0.128	6.698	0.000	0.606	1.107
trestbps	-0.0114	0.007	-1.613	0.107	-0.025	0.002
chol	-0.0124	0.003	-3.859	0.000	-0.019	-0.006
fbs	-0.3403	0.373	-0.912	0.362	-1.072	0.391
restecg	0.3310	0.247	1.342	0.180	-0.153	0.814
thalach	0.0390	0.006	6.062	0.000	0.026	0.052
exang	-0.6096	0.285	-2.142	0.032	-1.167	-0.052
oldpeak	-0.7124	0.154	-4.629	0.000	-1.014	-0.411
slope	0.6537	0.242	2.697	0.007	0.179	1.129
ca	-0.7788	0.131	-5.936	0.000	-1.036	-0.522
thal	-0.8299	0.200	-4.142	0.000	-1.223	-0.437

Dep. Variable:	target					
Method:	dydx					
At:	overall					
	dy/dx	std err	z	P> z 	[0.025	0.975]
age	0.0029	0.001	2.007	0.045	6.93e-05	0.006
sex	-0.2070	0.033	-6.267	0.000	-0.272	-0.142
cp	0.0925	0.012	7.823	0.000	0.069	0.116
trestbps	-0.0012	0.001	-1.621	0.105	-0.003	0.000
chol	-0.0013	0.000	-4.038	0.000	-0.002	-0.001
fbs	-0.0368	0.040	-0.914	0.361	-0.116	0.042
restecg	0.0358	0.027	1.348	0.178	-0.016	0.088
thalach	0.0042	0.001	6.782	0.000	0.003	0.005
exang	-0.0659	0.030	-2.172	0.030	-0.125	-0.006
oldpeak	-0.0770	0.015	-5.003	0.000	-0.107	-0.047
slope	0.0706	0.026	2.755	0.006	0.020	0.121
ca	-0.0842	0.013	-6.708	0.000	-0.109	-0.060
thal	-0.0897	0.020	-4.385	0.000	-0.130	-0.050

APPENDIX

A. *data_summary.py*

```

1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import seaborn as sns
5 from skimpy import skim
6
7
8 # read data
9 df = pd.read_csv('data/heart.csv')
10 skim(df)
11
12 X = df.drop('target', axis=1)
13 y = df['target']
14
15 corr = df.corr()
16
17
18 def get_lower_tri_heatmap(data, output="src/plots/correlation.png"):
19     mask = np.zeros_like(data, dtype=np.bool)
20     mask[np.triu_indices_from(mask)] = True
21
22     # Want diagonal elements as well
23     mask[np.diag_indices_from(mask)] = False
24
25     # Set up the matplotlib figure
26     f, ax = plt.subplots(figsize=(11, 9))
27
28     # Generate a custom diverging colormap
29     cmap = sns.diverging_palette(220, 10, as_cmap=True)
30
31     # Draw the heatmap with the mask and correct aspect ratio
32     sns_plot = sns.heatmap(data, mask=mask, cmap=cmap, vmax=.3, center=0,
33         square=True, linewidths=.5, cbar_kws={"shrink": .5})
34     # save to file
35     fig = sns_plot.get_figure()
36     fig.savefig(output)
37
38
39 def get_box_plots(data):
40     import plotnine as gg
41
42     for col in data:
43         plot = gg.ggplot(df) +\
44             gg.geom_boxplot(gg.aes(y=col, x='factor(target)', color='factor(target)'), notch
45 =True) +\
46             gg.xlab('target') +\
47             gg.coord_flip() +\
48             gg.scale_color_discrete(guide=False) +\
49             gg.theme(text=gg.element_text(size=24))
50         plot.save(f'plots/target-{col}.png', 'png')
51
52 if __name__ == '__main__':
53     # plot all sub-distributions
54     get_box_plots(X)
55
56     # corr plot
57     get_lower_tri_heatmap(corr)

```

B. associated_risk.py

```

1 import pandas as pd
2 from scipy.stats.contingency import relative_risk
3
4 from data_summary import X, y
5
6 # individual cross-tabs
7 age = pd.crosstab((X['age'] > 54).astype(int), y)
8 sex = pd.crosstab(X['sex'], y)
9 cp = pd.crosstab(X['cp'].isin([1, 2]).astype(int), y)
10 trestbps = pd.crosstab((X['trestbps'] > 130).astype(int), y)
11 chol = pd.crosstab((X['chol'] > 250).astype(int), y)
12 fbs = pd.crosstab(X['fbs'], y)
13 restecg = pd.crosstab(X['restecg'].isin([0, 1]).astype(int), y)
14 thalach = pd.crosstab((X['thalach'] > 150).astype(int), y)
15 exang = pd.crosstab(X['exang'], y)
16 oldpeak = pd.crosstab((X['oldpeak'] > 1.1).astype(int), y)
17 slope = pd.crosstab(X['slope'].isin([1, 2]).astype(int), y)
18 ca = pd.crosstab(X['ca'].isin([0, 1, 2]).astype(int), y)
19 thal = pd.crosstab(X['thal'].isin([2, 3]).astype(int), y)
20
21 crosstabs = [age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca,
22             thal]
23
24 # calculate ARD and RR
25 ard = lambda de, du, he, hu: round((de / (de + he)) - (du / (du + hu)), 4)
26 # rr = lambda de, du, he, hu: round((de / (de + he)) / (du / (du + hu)), 4)
27 risk_diff = dict()
28 risk_ratio = dict()
29 ci = dict()
30
31 for df in crosstabs:
32     De = df.iloc[1, 1]
33     Du = df.iloc[0, 1]
34     He = df.iloc[1, 0]
35     Hu = df.iloc[0, 0]
36
37     var = df.index.name
38     risk_diff[var] = ard(De, Du, He, Hu)
39     rr = relative_risk(De, De + He, Du, Du + Hu)
40     risk_ratio[var] = round(rr.relative_risk, 4)
41     ci[var] = rr.confidence_interval(confidence_level=0.95)
42
43 risk = pd.DataFrame.from_records([risk_diff, risk_ratio, ci],
44                                 index=['Risk Difference', 'Risk Ratio', 'RR 95% Confidence
45                                     Interval'])
46 risk = risk.transpose()
47 risk['RR 95% Confidence Interval'] = risk['RR 95% Confidence Interval'].apply(lambda row: tuple
48                                     ([round(i, 4) for i in row]))
49 # risk['Correlation'] = X.corrwith(y)
50 # risk = risk[['Correlation', 'Risk Difference', 'Risk Ratio']]
51 risk.set_axis(['age', 'sex', 'chest pain {1, 2}', 'resting blood pressure > 130', 'serum
52               cholestorl > 250 ml/dl',
53               'fasting blood sugar > 120 mg/dl', 'resting electrocardiographic results {0,
54               1}',
55               'maximum heart rate achieved > 150', 'exercise induced angina', 'oldpeak', '
56               slope',
57               'major vessels colored {0, 1, 2}', 'thal {2, 3}'], inplace=True)
58 risk.to_latex('src/risk-tables.tex', column_format='lrrrr')

```

C. *models.py*

```

1 import numpy as np
2 import pandas as pd
3 import statsmodels.api as sm
4 from sklearn.model_selection import train_test_split
5 from sklearn.metrics import classification_report, accuracy_score
6
7 from data_summary import X, y
8
9 # train/test split
10 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=42)
11
12 # logistic regression
13 log_reg = sm.Logit(y_train, X_train).fit()
14 with open('src/logistic-regression.tex', 'w') as f:
15     f.write(log_reg.summary().as_latex())
16
17 ame = log_reg.get_margeff(at='overall', method='dydx')
18 with open('src/marginal-effects.tex', 'w') as f:
19     f.write(ame.summary().as_latex())
20
21 odds = np.exp(log_reg.params)
22 print(odds)
23
24 print(log_reg.summary())
25 print(ame.summary())

```